Docket No.: PF-0419-2 DIV

Response Under 37 C.F.R. 1.116 - Expedited Procedure
Examining Group 1644

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop Appeal Brief-Patents, Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450 on September 8, 2003. By: Diane Kizer Printed: Diane Kizer

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re Application of: Bandman et al.

Title:	SH3-CONTAINING PROTEINS			
Serial No.:	09/925,122	Filing Date:	August 08, 2001	
Examiner:	Huynh, P.	Group Art Unit:	1644	

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

BRIEF ON APPEAL

Sir:

Further to the Notice of Appeal filed April 29, 2003, and received by the USPTO on May 6, 2003, herewith are three copies of Appellants' Brief on Appeal. Appellants hereby request a two-month extension of time in order to file this Brief. Authorized fees include the statutory fee of $410.00 for a two-month extension of time, as well as the $320.00 fee for the filing of this Brief.

This is an appeal from the decision of the Examiner finally rejecting claims 8, 45-46, 48 and 50-57 of the above-identified application.

09/12/2003 SDIRETAI 00000021 090108 09925122
01 FC:1402 320.00 DA

(1) REAL PARTY IN INTEREST

The above-identified application is assigned of record to Incyte Pharmaceuticals, Inc. (now Incyte Corporation, formerly known as Incyte Genomics, Inc.) (Reel 8997, Frame 0088), which is the real party in interest herein.

(2) RELATED APPEALS AND INTERFERENCES

Appellants, their legal representative and the assignee are not aware of any related appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

(3) STATUS OF THE CLAIMS

Claims rejected:	Claims 8, 45-46, 48 and 50-57
Claims allowed:	None
Claims canceled:	Claims 1-7 and 9-43
Claims withdrawn:	Claims 44, 47, 49, 58 and 59
Claims on Appeal:	Claims 8, 45-46, 48 and 50-57 (A copy of the claims on appeal, as amended, can be found in the attached Appendix).

(4) STATUS OF AMENDMENTS AFTER FINAL

An Amendment after Final Rejection under 37 C.F.R. § 1.116 is filed concurrently herewith. This Amendment removes issues for appeal. Therefore, it is believed that this Amendment will be entered.

(5) SUMMARY OF THE INVENTION

Appellants' invention is directed to antibodies which specifically bind to polypeptides, including SH3-containing proteins (HS3C), comprising the amino acid sequence of SEQ ID NO:1 (Specification, e.g., at page 3, lines 8-11; page 4, lines 6-7; page 15, lines 4-6 and 12-13; and page 30, line 28 to page 31, line 4). Appellants' invention also includes antibodies which specifically bind to

polypeptides which comprise naturally occurring variants of SEQ ID NO:1 (e.g., at page 31, lines 13-19). The invention further includes compositions comprising the foregoing antibodies (e.g., at page 35, lines 7-10), and methods of making the foregoing antibodies (e.g., at page 30, line 30 to page 32, line 16; and page 54, line 30 to page 55, line 16).

HS3C has strong chemical and structural homology with mouse formin binding protein 17 (FBP17) (GenBank ID 1255033; SEQ ID NO:5) (Specification, e.g., at page15 lines 17-20). In particular, HS3C and mouse FBP17 share 87% identity (e.g., at page 15, lines 19-20; and Figures 3A and 3B). In addition:

> "HS3C-1 is 265 amino acids in length and has various potential protein kinase phosphorylation sites for cAMP/cGMP dependent protein kinase at residue S_{141}, for casein kinase II at residues T_4, S_{13}, T_{42}, S_{165}, T_{191}, S_{214}, and S_{248}, for protein kinase C at residues T_{116} and T_{198}, and for tyrosine kinase at residues Y_{166} and Y_{244}. As shown in Figures 3A and 3B, HS3C-1 has chemical and structural homology with HS3C-2 (SEQ ID NO:3) and mouse FBP17 (GI 1255033; SEQ ID NO:5). In particular, HS3C-1 shares 51% and 87% identity with HS3C-2 and mouse FBP17, respectively. Several of the potential protein phosphorylation sites found in HS3C-1 are found in FB17 as well. HS3C-1, HS3C-2 and FBP17 share a similar SH3 domain between approximately residues C_{199} and Y_{250} in HS3C-1. The ALYTF sequence in FBP17 is shared by HS3C-1 and is similar in HS3C-2 (AIYHF). Residues F_{205}, W_{232}, and Y_{249} in HS3C-1, thought to be important in SH3 ligand binding, are shared by the other two proteins as well. HS3C-1 contains a distinctive leader sequence extending from M1 to approximately K39 that is not found in FBP17 and that may represent a signal peptide directing the protein to a particular cellular location. Northern analysis shows the expression of this sequence in various libraries, at least 48% of which are immortalized or cancerous and at least 48% of which involve inflammation or the immune response. Of particular note is the expression of HS3C-1 in prostate tissues associated with prostate tumors." (Specification at page 15, lines 13-30)

The antibodies of the present invention are useful, for example, for purifying and detecting polypeptides which have specific uses in toxicology testing, drug discovery, and disease diagnosis (Specification, e.g., at page 26, lines 9-17; page 38, line 27 to page 39, line 13; and page 46, lines 12-15).

(6) ISSUES

1. Whether claims 8 and 50-52 are obvious under 35 U.S.C. § 103(a) over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, page 93).

2. Whether claims 45, 46 and 48 are obvious under 35 U.S.C. § 103(a) over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pages 319-356 and 626-629).

3. Whether claim 45 is obvious under 35 U.S.C. § 103(a) over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of US Pat No. 4, 946,778 (August 1990).

4. Whether claims 45, 56 and 57 are obvious under 35 U.S.C. § 103(a) over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of US Pat No. 6,180,370B (filed June 1995).

5. Whether claims 53-55 are obvious under 35 U.S.C. § 103(a) over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pages 139-149).

6. Whether claims 50-57 meet the written description requirement of 35 U.S.C. §112, first paragraph.

7. Whether claims 8, 45, 46, 48 and 50-57 meet the enablement requirement of 35 U.S.C. §112, first paragraph.

(7) GROUPING OF THE CLAIMS

As to Issue 1

Claims 8 and 50-52 are grouped together.

As to Issue 2

Claims 45, 46 and 48 are grouped together.

As to Issue 3

Claim 45 is the only claim on appeal under this issue.

As to Issue 4

Claims 45, 56 and 57 are grouped together.

As to Issue 5

Claims 53-55 are grouped together.

As to Issue 6

Claims 50-57 are grouped together.

As to Issue 7

Claims 8, 45, 46, 48 and 50-57 are grouped together.

(8) APPELLANTS' ARGUMENTS

Issue 1--Whether claims 8 and 50-52 are obvious under 35 U.S.C. § 103(a) over Chan et al. in view of Harlow et al.

Claims 8 and 50-52 were rejected under 35 U.S.C. § 103(a) because the claimed antibodies are allegedly obvious over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, page 93). This rejection is traversed.

The rejection is based on the allegation that Chan et al. teach a biologically active or immunogenic fragment that is within the scope of the claimed antibodies. The Examiner has mischaracterized the claimed antibodies as allegedly binding to "a biologically active fragment or an immunogenic fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1 or to a biologically active fragment or an immunogenic fragment of a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1." (Final Office Action at page 7.) Independent claim 8 recites antibodies which **specifically** bind to a polypeptide comprising

SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Chan et al. provides no recognition of such a sequence. Moreover, Chan et al. does not teach antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Indeed, Chan et al. does not teach antibodies at all.

The biologically active fragment taught in Figure 3A of Chan et al., designated FBP17, is a 53 amino acid fragment that is neither a polypeptide comprising SEQ ID NO:1, nor a naturally occurring variant thereof which is ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. The entire FBP17 sequence (GenBank accession number AAC52479; GI1255033) is 237 amino acids in length. A CLUSTALW alignment of the 237 amino acid FBP17 sequence with the 265 amino acid sequence of SEQ ID NO:1 indicates a sequence identity of 78% (207/265 amino acids) between the two sequences (See Attachment 1). Without a recognition of the amino acid sequence of SEQ ID NO:1 or a polypeptide at least 90% identical to SEQ ID NO:1, one would not be able to make an antibody which specifically binds the recited proteins.

To support an obviousness rejection under 35 U.S.C. § 103, "all the claim limitations must be taught or suggested by the prior art." M.P.E.P. § 2143.03. In addition, "the reference teachings must somehow be modified in order to meet the claims. The modification must be one which would have been obvious to one of ordinary skill in the art at the time the invention was made." M.P.E.P. § 706.02. Since the claim language distinguishes the recited antibodies from the teachings of Chan et al., the Examiner has not convincingly shown how the teachings of Chan et al. and/or Harlow et al. could be modified in order to arrive at the claimed subject matter. Therefore, the Examiner has not met the requirements for a *prima facie* showing of obviousness under 35 U.S.C. § 103(a).

Therefore, this rejection of claims 8 and 50-52 should be overturned.

Issue 2--Whether claims 45, 46 and 48 are obvious under 35 U.S.C. § 103(a) over Chan et al. in view of Harlow et al.

Claims 45, 46 and 48 were rejected under 35 U.S.C. § 103(a) because the claimed antibodies are allegedly obvious over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pages 319-356 and 626-629). This rejection is traversed.

The rejection is based on the allegation that Chan et al. teach a biologically active or immunogenic fragment that is within the scope of the claimed antibodies. The Examiner has mischaracterized the claimed antibodies as allegedly binding to "a biologically active fragment or an immunogenic fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1 or to a biologically active fragment or an immunogenic fragment of a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1." (Final Office Action at page 7.) As discussed above under Issue 1, the claims recite antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Chan et al. provides no recognition of such a sequence. Moreover, Chan et al. does not teach antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Indeed, Chan et al. does not teach antibodies at all.

To support an obviousness rejection under 35 U.S.C. § 103, "all the claim limitations must be taught or suggested by the prior art." M.P.E.P. § 2143.03. In addition, "the reference teachings must somehow be modified in order to meet the claims. The modification must be one which would have been obvious to one of ordinary skill in the art at the time the invention was made." M.P.E.P. § 706.02. Since the claim language distinguishes the recited antibodies from the teachings of Chan et al., the Examiner has not convincingly shown how the teachings of Chan et al. and/or Harlow et al. could be modified in order to arrive at the claimed subject matter. Therefore, the Examiner has not met the requirements for a *prima facie* showing of obviousness under 35 U.S.C. § 103(a).

Therefore, this rejection of claims 45, 46 and 48 should be overturned.

Issue 3--Whether claim 45 is obvious under 35 U.S.C. § 103(a) over Chan et al. in view of US Pat No. 4, 946,778

Claim 45 was rejected under 35 U.S.C. § 103(a) because the claimed antibodies are allegedly obvious over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of US Pat No. 4,946,778 (Ladner et al.). This rejection is traversed.

The rejection is based on the allegation that Chan et al. teach a biologically active or immunogenic fragment that is within the scope of the claimed antibodies. The Examiner has mischaracterized the claimed antibodies as allegedly binding to "a biologically active fragment or an immunogenic fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1 or to a biologically active fragment or an immunogenic fragment of a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1." (Final Office Action at page 7.) As discussed above under Issues 1-2, the claims recite antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Chan et al. provides no recognition of such a sequence. Moreover, Chan et al. does not teach antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Indeed, Chan et al. does not teach antibodies at all.

To support an obviousness rejection under 35 U.S.C. § 103, "all the claim limitations must be taught or suggested by the prior art." M.P.E.P. § 2143.03. In addition, "the reference teachings must somehow be modified in order to meet the claims. The modification must be one which would have been obvious to one of ordinary skill in the art at the time the invention was made." M.P.E.P. § 706.02. Since the claim language distinguishes the recited antibodies from the teachings of Chan et al., the Examiner has not convincingly shown how the teachings of Chan et al. and/or Ladner et al. could be modified in order to arrive at the claimed subject matter. Therefore, the Examiner has not met the requirements for a *prima facie* showing of obviousness under 35 U.S.C. § 103(a) and this rejection of claim 45 should be overturned.

Issue 4--Whether claims 45, 56 and 57 are obvious under 35 U.S.C. § 103(a) over Chan et al. in view of US Pat No. 6,180,370B

Claims 45, 56 and 57 were rejected under 35 U.S.C. § 103(a) because the claimed antibodies are allegedly obvious over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of US Pat No. 6,180,370B (Queen et al.). This rejection is traversed.

The rejection is based on the allegation that Chan et al. teach a biologically active or immunogenic fragment that is within the scope of the claimed antibodies. The Examiner has mischaracterized the claimed antibodies as allegedly binding to "a biologically active fragment or an immunogenic fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1 or to a biologically active fragment or an immunogenic fragment of a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1." (Final Office Action at page 7.) As discussed above under Issues 1-3, the claims recite antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Chan et al. provides no recognition of such a sequence. Moreover, Chan et al. does not teach antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Indeed, Chan et al. does not teach antibodies at all.

To support an obviousness rejection under 35 U.S.C. § 103, "all the claim limitations must be taught or suggested by the prior art." M.P.E.P. § 2143.03. In addition, "the reference teachings must somehow be modified in order to meet the claims. The modification must be one which would have been obvious to one of ordinary skill in the art at the time the invention was made." M.P.E.P. § 706.02. Since the claim language distinguishes the recited antibodies from the teachings of Chan et al., the Examiner has not convincingly shown how the teachings of Chan et al. and/or Queen et al. could be modified in order to arrive at the claimed subject matter. Therefore, the Examiner has not met the requirements for a *prima facie* showing of obviousness under 35 U.S.C. § 103(a) and this rejection of claims 45, 56 and 57 should be overturned.

Issue 5--Whether claims 53-55 are obvious under 35 U.S.C. § 103(a) over Chan et al. in view of Harlow et al.

Claims 53-55 were rejected under 35 U.S.C. § 103(a) because the claimed antibodies are allegedly obvious over Chan et al. (EMBO Journal, 1996, 15:1045-1054) in view of Harlow et al. (in Antibodies a Laboratory Manual, 1988, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pages 139-149). This rejection is traversed.

The rejection is based on the allegation that Chan et al. teach a biologically active or immunogenic fragment that is within the scope of the claimed antibodies. The Examiner has mischaracterized the claimed antibodies as allegedly binding to "a biologically active fragment or an immunogenic fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1 or to a biologically active fragment or an immunogenic fragment of a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1." (Final Office Action at page 7.) As discussed above under Issues 1-4, the claims recite antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Chan et al. provides no recognition of such a sequence. Moreover, Chan et al. does not teach antibodies which **specifically** bind to a polypeptide comprising SEQ ID NO:1, or a polypeptide comprising a naturally occurring variant thereof, the variant being ***at least 90% identical to SEQ ID NO:1*** and having ***HS3C activity***. Indeed, Chan et al. does not teach antibodies at all.

To support an obviousness rejection under 35 U.S.C. § 103, "all the claim limitations must be taught or suggested by the prior art." M.P.E.P. § 2143.03. In addition, "the reference teachings must somehow be modified in order to meet the claims. The modification must be one which would have been obvious to one of ordinary skill in the art at the time the invention was made." M.P.E.P. § 706.02. Since the claim language distinguishes the recited antibodies from the teachings of Chan et al., the Examiner has not convincingly shown how the teachings of Chan et al. and/or Harlow et al. could be modified in order to arrive at the claimed subject matter. Therefore, the Examiner has not met the requirements for a *prima facie* showing of obviousness under 35 U.S.C. § 103(a).

Therefore, this rejection of claims 53-55 should be overturned.

Issue 6--Whether claims 50-57 meet the written description requirement of 35 U.S.C. §112, first paragraph.

Claims 50-57 stand rejected under 35 U.S.C. § 112, first paragraph, based on the allegation that the Specification does not describe the subject matter in such a way as to reasonably convey to one of skill in the art that the inventors, at the time the application was filed, had possession of the claimed invention. The Examiner asserts that "[g]iven the lack of any immunogenic polypeptide fragment and additional representative species of polypeptide other than the polypeptide of SEQ ID NO:1 and 3 to which the antibody binds wherein the antibody is polyclonal, monoclonal, chimeric, humanized, Fab fragment, F(ab')2 fragment thereof, one of skill in the art would reasonably conclude that the disclosure fails to provide a representative number of species to describe the genus" (Final Office Action at pp.5-6). This rejection is traversed.

In making this rejection, the Examiner assails Appellants' use of the term "having" in claims 50 and 53 (i.e., "...immunizing an animal with a polypeptide ***having*** an amino acid sequence of SEQ ID NO:1..." [emphasis added]) by stating that "there is insufficient written description about the structure associated with function of a method of making polyclonal or monoclonal antibody [*sic*] by immunizing an animal with a polypeptide "having" an amino acid sequence of an immunogenic fragment of SEQ ID NO:1 for in vivo treatment of any disease or for diagnostic assays. The term "having" is open-ended. It expands the immunogenic fragment to include additional amino acids at either end. There is inadequate written description about the additional undisclosed amino acids to be added to the immunogenic fragment." (Final Office Action at page 5.) First note that the Examiner has mischaracterized claims 50-57 as they do not recite "immunizing an animal with a polypeptide "having" an amino acid sequence of an immunogenic fragment of SEQ ID NO:1 for in vivo treatment of any disease or for diagnostic assays."

The Examiner's assertions seem to imply that the use of the transitional phrase "having" in claims 50 and 53 requires that the Specification provide a written description of any possible element which could be part of, but is not essential to, the claimed subject matter. However the transitional phrase "'[c]omprising' [or having] is a term of art used in claim language which means that the named elements are essential, but other elements may be added and still form a construct within the scope of

the claim." M.P.E.P. § 2111.03 (citing *Genentech, Inc. v. Chiron Corp.*, 112 F.3d 495, 501, 42 USPQ2d 1608, 1613 (Fed. Cir. 1997)). The Specification has described numerous examples of polypeptides "having" or "comprising" the recited polypeptides and variants of SEQ ID NO:1, such as fusion proteins and coupled proteins (Specification, e.g., at page 9, lines 3-9; page 21, lines 4-7; page 27, lines 7-22; page 31, lines 13-19; ;and page 55, lines 9-16). One of skill in the art would understand that Appellants had possession of the described polypeptides, "having" or "comprising" the recited polypeptides and variants of SEQ ID NO:1, without an explicit disclosure of every possible element which could be a part of, but is not essential to, the claimed subject matter.

Moreover, the claims recite antibodies which specifically bind to epitopes on the recited polypeptides and variants of SEQ ID NO:1. For example, the claimed antibodies specifically bind to "an epitope of a polypeptide of SEQ ID NO:1," and "an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." Since the claimed antibodies specifically bind to the recited epitopes, and since there is an adequate written description of the recited polypeptides and variants of SEQ ID NO:1, any additional amino acid residues "at either end" of the recited polypeptides and variants of SEQ ID NO:1 are not essential to the claimed subject matter. Therefore, it is irrelevant whether there is a detailed written description of additional amino acid residues "at either end" of the recited polypeptides and variants of SEQ ID NO:1.

The Examiner further asserts that "[g]iven the lack of any immunogenic polypeptide fragment and additional representative species of polypeptide other than the polypeptide of SEQ ID NO:1...to which the antibody binds...one of skill in the art would reasonably conclude that the disclosure fails to provide a representative number of species to describe the genus" (Final Office Action at pp.5-6). To the contrary. Claim 8, for example, recites antibodies which specifically bind to "an epitope of a polypeptide of SEQ ID NO:1," and "an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." One of ordinary skill in the art would recognize polypeptide sequences which are possible epitopes of SEQ ID NO:1. The amino acid sequence of SEQ ID NO:1 provides the necessary framework for the recited epitopes. To recite every possible epitope would needlessly

clutter the application. It would be routine for one of skill in the art to determine whether any possible epitope had immunogenic activity, based on the methods recited in the Specification at, for example page 9, lines 27-30; page 30, line 27 to page 32, line 23; and page 55, lines 1-16. Accordingly, the Specification provides an adequate written description of the claimed antibodies which specifically bind to the recited epitopes.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. § 112, first paragraph, are well established by case law.

> . . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*. The invention is, for purposes of the "written description" inquiry, *whatever is now claimed.* *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

> . . . Mention of representative compounds encompassed by generic claim language ***clearly is not required by Section 112 or any other provision of the statute***. But, where no explicit description of a generic invention is to be found in the specification...mention of representative compounds may provide an implicit description upon which to base generic claim language. *In re Robins*, 429 F.2d 452, 456-57, 166 USPQ 552, 555 (CCPA 1970) [emphasis added]

> . . . [I]t has been consistently held that the naming of one member of such a group is not, in itself, a proper basis for a claim to the entire group. However, ***it may not be necessary to enumerate a plurality of species if a genus is sufficiently identified in an application by 'other appropriate language.'*** *In re Grimme*, 274 F.2d 949, 952, 124 USPQ 499, 501 (CCPA 1960) [emphasis added]

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that:

> An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such

> characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met. [footnotes omitted]

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

A. The Specification provides an adequate written description of the claimed antibodies which specifically bind to the recited "variants" of SEQ ID NO:1.

The subject matter encompassed by claims 50-57 is either disclosed by the Specification or is conventional or well known to one skilled in the art.

First note that the "variant" language of independent claim 8, from which claims 50-57 depend, recites polypeptides "having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." The polypeptide sequence of SEQ ID NO:1 is explicitly disclosed in the Specification. See, for example, the Sequence Listing and Figures 1A, 1B, 1C, 1D, 3A and 3B. Variants of SEQ ID NO:1 are described in the Specification at, for example, page 7, line 30 to page 8, line 15; page 8, lines 19-22; page 14, line 23 to page 15, line 1; and page 16, lines 21-26. In addition, a specific assay to measure HS3C activity is disclosed in the Specification at, for example, page 54, lines 18-28.

One of ordinary skill in the art would recognize polypeptide sequences which are variants that are at least 90% identical to SEQ ID NO:1. Given any naturally occurring polypeptide sequence, it would be routine for one of skill in the art to recognize whether it was a variant of SEQ ID NO:1. It would also be routine to determine whether such a variant had HS3C activity, using the disclosed HS3C binding assay. Accordingly, the Specification provides an adequate written description of the claimed antibodies which specifically bind to the recited polypeptide variants of SEQ ID NO:1.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue (which are hence relevant to claims to proteins encoded by the DNA and to antibodies which specifically bind those proteins) commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. § 112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polypeptides bound by the claimed antibodies in terms of chemical structure, rather than functional characteristics. For example, the language of independent claim 8 (as amended by the Amendment After Final filed herewith) recites chemical structure to define the claimed genus:

> 8. An isolated antibody selected from the group consisting of:
> a) an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1 or SEQ ID NO:3, wherein the antibody binds to an epitope of a polypeptide of SEQ ID NO:1 or SEQ ID NO:3, and
> b) an antibody which specifically binds to a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1 or SEQ ID NO:3, wherein the antibody binds to an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1 or SEQ ID NO:3, said naturally occurring amino acid sequence having HS3C activity.

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polypeptides specifically bound by the claimed antibodies. The polypeptides defined by the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base the written description inquiry "on whatever is now claimed," the Examiner failed to provide an appropriate

analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

The Patent Office Guidelines indicate that evidence that Appellants were in possession of the claimed invention can include "complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, **or** some combination of such characteristics" (P.T.O. Guidelines, *supra*; emphasis added). The claimed antibodies which specifically bind the recited variants of the SEQ ID NO:1 polypeptide have been described by chemical structure (e.g., relation of the recited polypeptide variants to SEQ ID NO:1), physical properties (e.g., occurrence in nature of the recited polypeptide variants), and chemical properties (e.g., possession of HS3C activity by the recited polypeptide variants; specific binding of the claimed antibodies to the recited polypeptide variants). Therefore, the written description requirement has been met.

2. The present claims do not define a genus which is "highly variant"

Furthermore, the claims at issue do not describe a genus which could be characterized as "highly variant." Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Board's attention is directed to the reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078) (See Attachment 2). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to SH3-containing proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as SH3-containing proteins and which have as little as 30% identity over at least 150 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, antibodies which specifically bind to "a polypeptide comprising a naturally occurring amino

acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1... wherein the antibody binds to an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1...said naturally occurring amino acid sequence having HS3C activity" (note that SEQ ID NO:1 has 265 amino acid residues). This variation is far less than that of all potential SH3-containing proteins related to SEQ ID NO:1, i.e., those SH3-containing proteins having as little as 30% identity over at least 150 residues to SEQ ID NO:1.

3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. § 112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those cases was based on the state of the art at essentially the "dark ages" of recombinant DNA technology.

The present application has a priority date of November 13, 1997. Much has happened in the development of recombinant DNA technology in the 20 or so years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances, one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed antibodies which specifically bind the recited polypeptide variants at the time of filing of this application.

4. The Examiner has attempted to apply a standard for written description different from that which is required by law

The Examiner has alleged that claims 50-57 do not comply with the requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph because:

- Given the lack of any immunogenic polypeptide fragment and additional representative species of polypeptide other than the polypeptide of SEQ ID

> NO:1...to which the antibody binds wherein the antibody is polyclonal, monoclonal, chimeric, humanized, Fab fragment, F(ab')2 fragment thereof, one of skill in the art would reasonably conclude that the disclosure fails to provide a representative number of species to describe the genus. Thus, Applicant was not in possession of the claimed genus. (Final Office Action at pages 5-6.)

Appellants submit that neither the written description requirement of 35 U.S.C. 112, first paragraph nor any case law that interprets the statute has ever set forth such a standard. Furthermore, case law in the area of the written description requirement of 35 U.S.C. 112, first paragraph is clear with regard to the details considered sufficient to describe a claimed genus:

> . . . Mention of representative compounds encompassed by generic claim language ***clearly is not required by Section 112 or any other provision of the statute***. But, where no explicit description of a generic invention is to be found in the specification...mention of representative compounds may provide an implicit description upon which to base generic claim language. *In re Robins*, 429 F.2d 452, 456-57, 166 USPQ 552, 555 (CCPA 1970) [emphasis added]

> . . . [I]t has been consistently held that the naming of one member of such a group is not, in itself, a proper basis for a claim to the entire group. However, ***it may not be necessary to enumerate a plurality of species if a genus is sufficiently identified in an application by 'other appropriate language.'*** *In re Grimme*, 274 F.2d 949, 952, 124 USPQ 499, 501 (CCPA 1960) [emphasis added]

The Specification sets forth a description of the claimed polypeptide variants using "other appropriate language" as indicated above in connection with the remarks regarding "a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." The claimed variants have been described in terms of their relationship to the chemical structure of SEQ ID NO:1 and structural requirements for biological and immunological activity at, for example, pp. 57-58 of the Sequence Listing; Figures 1A, 1B, 1C, 1D, 3A and 3B; page 7, line 30 to page page 8, line 15; page 8, lines 19-22; page 14, line 23 to page 15, line 1; and page 16, lines 21-26. The Specification provides a means of identifying naturally occurring functional variants having 90% sequence identity with SEQ ID NO:1 and having HS3C activity at, for example, p. 14, line 23 to p. 15, line 1; p. 39, line 21 to p. 40, line 3; Example VI at p. 52; and Example X at p. 54. Appellants therefore submit that the "genus is sufficiently identified in [the instant] application by 'other appropriate language'" as stated in *In re Grimme*, 274 F.2d 949,

952, 124 USPQ 499, 501 (CCPA 1960). Furthermore, Appellants submit that "a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing" as stated in the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001. Accordingly, claims 50-57 meet the statutory requirements for written description under 35 U.S.C. 112, first paragraph.

5. Summary

The Examiner failed to base the written description inquiry "on whatever is now claimed." Consequently, the Examiner did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polypeptides recited by the present claims is adequately described, as evidenced by Brenner et al. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Examiner.

For at least the reasons set forth above, the Specification provides an adequate written description of the claimed antibodies which specifically bind to the recited polypeptide "variants," and this rejection should be overturned.

Issue 7--Whether claims 8, 45, 46, 48 and 50-57 meet the enablement requirement of 35 U.S.C. §112, first paragraph.

Claims 8, 45, 46, 48 and 50-57 stand rejected under 35 U.S.C. § 112, first paragraph based on the allegation that the Specification does not describe the subject matter of the invention in such a way as to enable one of skill in the art to make and/or use antibodies which specifically bind to the recited "variants" of SEQ ID NO:1. In particular, the Examiner asserts that the Specification "**does not** reasonably provide ennoblement [*sic*] for (1) *any* isolated antibody...which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to *any*

"epitope" of a polypeptide of SEQ ID NO:1, and b) *any* antibody which specifically binds to a polypeptide comprising *any* "naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to *any* "epitope" of *any* "polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid having HS3C activity..." (Final Office Action at page 15; emphasis in original). Such, however, is not the case.

The Specification discloses methods to make antibodies which specifically bind to a polypeptide having **any** particular amino acid sequence (e.g., at page 30, line 27 to page 32, line 16; and page 55, lines 1-16). Given the information provided by SEQ ID NO:1 (the amino acid sequence of HS3C), one of skill in the art would be able to routinely obtain antibodies which specifically bind to any of the recited polypeptides and variants of SEQ ID NO:1, including "an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to an epitope of a polypeptide of SEQ ID NO:1," and "an antibody which specifically binds to a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." For example, an animal could be immunized with any of the recited polypeptides, variants and fragments of SEQ ID NO:1, antibodies could be isolated from the animal, and the antibodies could be screened to identify antibodies which specifically bind to the polypeptide or variant.

Likewise, the specification discloses methods to use antibodies which specifically bind to a polypeptide having **any** particular amino acid sequence in, for example, the purification of such polypeptides (e.g., at page 55, lines 18-28), the detection and/or measurement of such polypeptides (e.g., at page 26, lines 9-17; and page 38, line 26 to page 39, line 13), and the competitive screening of drug candidates (e.g., at page 46, lines 12-15). Given the information provided by SEQ ID NO:1 (the amino acid sequence of HS3C), one of skill in the art would be able to routinely use antibodies which specifically bind to any of the recited polypeptides and variants of SEQ ID NO:1, including "an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to an epitope of a polypeptide of SEQ ID NO:1," and "an antibody which specifically binds to a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, wherein the antibody binds to an epitope

of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino acid sequence having HS3C activity." For example, an antibody which specifically binds to any of the recited polypeptides and variants of SEQ ID NO:1 could be coupled to an activated chromatographic resin, and this resin could then be used in an immunoaffinity column to purify the polypeptide or variant.

In making this rejection the Examiner assails Appellants' use of the term "comprising" by stating that "the term comprising or having is open-ended. It expands the "immunogenic fragment" to include additional amino acids at either or both ends. There is insufficient guidance as to the undisclosed amino acid added to the immunogenic fragment for making antibody that binds to the full-length polypeptide of SEQ ID NO:1 or any polypeptide having 10% difference in the amino acid sequence of SEQ ID NO:1, which is equivalent to having 26-27 amino acid difference in SEQ ID NO:1, or any epitope of said polypeptides." (Final Office Action at page 19.) To the contrary, the claimed antibodies are fully enabled by the Specification.

The Examiner's assertions seem to imply that the use of the transitional phrase "comprising" in claim 8 (and therefore claims 45, 46, 48 and 50-57 which depend therefrom) requires that the Specification provide enablement for any possible element which could be a part of, but is not essential to, the claimed subject matter. However, the transitional phrase "'[c]omprising' is a term of art used in claim language which means that the named elements are essential, but other elements may be added and still form a construct within the scope of the claim." M.P.E.P. § 2111.03 (citing *Genentech, Inc. v. Chiron Corp.*, 112 F.3d 495, 501, 42 USPQ2d 1608, 1613 (Fed. Cir. 1997)). The Specification has disclosed numerous examples of polypeptides "comprising" the recited polypeptides, variants, and fragments of SEQ ID NO:1, such as fusion proteins and coupled proteins (Specification, e.g., at page 9, lines 3-9; page 21, lines 4-7; page 27, lines 7-22; page 31, lines 13-19; ;and page 55, lines 9-16). One of skill in the art would understand how to make and use antibodies which specifically bind to the disclosed polypeptides, "comprising" the recited polypeptides and variants of SEQ ID NO:1 without an explicit disclosure of every possible element which could be a part of, but is not essential to, the claimed subject matter.

Moreover, the claims recite antibodies which specifically bind to epitopes on the recited polypeptides and variants of SEQ ID NO:1. For example, the claimed antibodies specifically bind to "an epitope of a polypeptide of SEQ ID NO:1" and "an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1, said naturally occurring amino

acid sequence having HS3C activity." Since the claimed antibodies specifically bind to the recited epitopes, and since a skilled artisan would know how to make and use antibodies which specifically bind to epitopes of the recited polypeptides and variants of SEQ ID NO:1, any additional amino acid residues at "either or both ends" of the recited polypeptides and variants of SEQ ID NO:1 are not essential to the claimed subject matter. Therefore, it is irrelevant whether the Specification describes antibodies which specifically bind to additional amino acid residues at "either or both ends" of the recited polypeptides and variants of SEQ ID NO:1.

With regard to the Examiner's concern over "any polypeptide having 10% difference in the amino acid sequence of SEQ ID NO:1, which is equivalent to having 26-27 amino acid difference in SEQ ID NO:1, or any epitope of said polypeptides" (Final Office Action at page 19), Appellants submit that this concern is irrelevant to the enablement of the claimed antibodies. Antibodies which specifically bind to a polypeptide can be made as long as that polypeptide, or fragments thereof, are available; there is no restriction on the amino acid sequence of polypeptides that can be used to make antibodies. Since a polypeptide having **any** amino acid sequence (including any amino acid sequence that is at least 90% identical to SEQ ID NO:1 and any naturally occurring amino acid sequence that is at least 90% identical to SEQ ID NO:1) can be used to make antibodies using the methods disclosed in the Specification, it is not necessary to identify particular naturally occurring amino acid sequences that are at least 90% identical to SEQ ID NO:1 that could be used in this manner. Moreover, the question of whether a polypeptide or variant thereof has biological function (see item (2) on page 22 of the Final Office Action) is irrelevant to the enablement of antibodies which specifically bind to that polypeptide or variants thereof. Even if a polypeptide variant has no known biological function, it can nevertheless be used to make antibodies which specifically bind to that polypeptide variant without undue experimentation.

In further support of this rejection, the Examiner states that the enablement rejection is maintained because "there is sufficient [*sic*] guidance as the antigenic determinant (i.e. the specific amino acid sequence of the immunogen or polypeptide fragment) used by applicant to make any antibody mentioned above that binds to *any* "epitope" of a polypeptide of SEQ ID NO:1 or *any* "epitope" of a polypeptide having 10% difference (90% identity) in the amino acid sequence of SEQ ID NO:1..." (Final Office Action at page 22; emphasis in original). However, no such explicit guidance would be required as one of skill in the art would know how to choose the "amino acid sequence of the immunogen or polypeptide fragment" used to make the claimed antibodies (e.g., Specification at page

55, lines 1-16). All that is necessary to satisfy the enablement requirement of 35 U.S.C. § 112, first paragraph, is that a skilled artisan would reasonably understand how to make and use the claimed antibodies, without undue experimentation.

Antibodies which specifically bind to a polypeptide can be made as long as that polypeptide, or fragments thereof, are available; there is no restriction on the amino acid sequence of polypeptides that can be used to make antibodies. Thus, any of the recited polypeptides (including polypeptides having an amino acid sequence that is at least 90% identical to SEQ ID NO:1, a naturally occurring amino acid sequence that is at least 90% identical to SEQ ID NO:1, or an amino acid sequence that is a fragment of SEQ ID NO:1) can be used to make antibodies using the methods disclosed in the specification. A skilled artisan would be able to routinely determine whether a given antibody produced in this way specifically binds to the polypeptide used to make that antibody, without undue experimentation. For example, an antibody can be subjected to a screen which identifies antibodies that specifically bind to the polypeptide.

Citing Kuby et al. (Immunology, Second edition, 1994, W.H. Freeman and Company, New York, NY, page 94), the Examiner states that "[i]mmunization with a peptide fragment may result in antibody specificity that differs from the **antibody specificity** directed against the native full-length polypeptide." (Final Office Action at pages 19-20; emphasis in original). In making this statement, the Examiner has improperly imported additional limitations into the claims. In making an antibody which specifically binds to the full-length polypeptide of SEQ ID NO:1, there is no requirement to use a peptide fragment of SEQ ID NO:1. In fact, there is no requirement that **any particular polypeptide** be used to make any of the claimed antibodies. The claims recite antibodies which specifically bind to the recited polypeptides and variants of SEQ ID NO:1. All that is necessary to satisfy the enablement requirement of 35 U.S.C. § 112, first paragraph, is that a skilled artisan would reasonably understand how to make and use the claimed antibodies, without undue experimentation.

As stated above, antibodies which specifically bind to a polypeptide can be made as long as that polypeptide, or fragments thereof, are available; there is no restriction on the amino acid sequence of polypeptides that can be used to make antibodies. Thus, the full-length polypeptide of SEQ ID NO:1 can be used to make antibodies using the methods disclosed in the specification. A skilled artisan would be able to routinely determine whether a given antibody produced in this way specifically binds to the full-length polypeptide of SEQ ID NO:1, without undue experimentation. For example, an

antibody can be subjected to a screen which identifies antibodies that specifically bind to the full-length polypeptide of SEQ ID NO:1.

The Examiner attempts to provide further support for this rejection by citing Ngo et al. (in The Protein Folding Problem and Tertiary Structure Prediction, 1994, Birkhauser Boston, pages 492-495), Abaza et al. (J. Prot. Chem., 1992, 11:433-444), and Skolnick et al. (Trends in Biotech., 2000, 18:34-39). The Examiner asserts that "Ngo *et al* teach that the amino acid positions within the polypeptide/protein that can tolerate change such as conservative substitution or no substitution, addition or deletion which are critical to maintain the protein's structure/function will require guidance" (Final Office Action at page 20). Once again, the question is not whether the recited polypeptides which are specifically bound by the claimed antibodies retain the structure and/or function of the SEQ ID NO:1 polypeptide. The relevant question, for the purposes of enablement under 35 U.S.C. § 112, first paragraph, is whether a skilled artisan could make and use the claimed antibodies which specifically bind the recited polypeptides. Regardless of whether a variant or fragment of SEQ ID NO:1 maintains the structure and/or function of the SEQ ID NO:1 polypeptide, that variant or fragment could still be used to make antibodies, without undue experimentation. Thus, the enablement requirement is satisfied.

With respect to the Abaza reference, the Examiner contends that "even a single amino acid substitution outside the antigenic site can exert drastic effects on the reactivity of a protein with monoclonal antibody against the site" (Final Office Action at page 20). However, whether or not the Examiner's contention is true, it has no bearing on whether one of skill in the art could make and/or use the claimed antibodies, without undue experimentation. Even if a polypeptide variant has an amino acid substitution which drastically affects the reactivity of a monoclonal antibody which specifically binds to the parent polypeptide, a skilled artisan would still know how to use that polypeptide variant to make antibodies which specifically bind to the polypeptide variant. In addition, one of skill in the art would know how to use such antibodies to purify and/or detect the polypeptide variant. Therefore, the claimed antibodies meet the enablement requirement of 35 U.S.C. § 112, first paragraph.

With respect to the Skolnick article, the Examiner asserts that "sequence-based methods for function prediction are inadequate and knowing a protein's structure does not tell one its function (Final Office Action at page 20). Again, the question of whether a polypeptide or variant thereof has any particular biological function is irrelevant to the enablement of antibodies which specifically bind to that polypeptide or variants thereof. Even if a polypeptide or variant thereof has no known biological

function, it can nevertheless be used to make antibodies which specifically bind to that polypeptide or variant without undue experimentation.

With respect to polyclonal or monoclonal antibodies which specifically bind to the recited polypeptide sequences of claim 8 and an acceptable excipient or suitable carrier, the Examiner states that "the specification fails to provide any *in vivo* working examples, or guidance with respect to treating a patient suffering from any specific disease or conditions that may or may [*sic*] associated with the expression of "HS3C" (Final Office Action at pages 20-21). In addition, the Examiner states that "[t]here are no working examples in the specification as filed that the claimed antibody ever been made [*sic*]" (Final Office Action at page 23). By these statements, it appears that the Examiner would require an actual reduction to practice of the claimed invention in order to satisfy the enablement requirement of 35 U.S.C. § 112, first paragraph. However, an actual reduction to practice is not necessary.

There is no legal requirement that an invention actually be reduced to practice in order for that invention to be patentable. The amino acid sequence of the polypeptide of SEQ ID NO:1 has been explicitly disclosed in the specification (see, e.g., the Sequence Listing and Figures 1A, 1B, 1C, 1D, 3A and 3B). Methods of making and using antibodies which specifically bind to polypeptides (including polypeptides based on the SEQ ID NO:1 polypeptide) have also been disclosed in the specification (e.g., at page 30, line 27 to page 32, line 16; and page 55, lines 1-28). In conjunction with the disclosure in the specification and the knowledge in the art at the time the application was filed, a skilled artisan would know how to make and use the claimed antibodies. Thus, the constructive reduction to practice of the claimed antibodies more than adequately provides enablement for the claimed invention.

The Examiner asserts, with respect to chimeric antibodies, that "[i]n the absence of in vivo working examples, it is unpredictable for the following reasons: (1) the antibody may be inactivated before producing an effect, i.e. such as inherently short half-life of the antibody; (2) the antibody may not reach the target area; and (3) other function properties, known or unknown, may make the antibody unsuitable for *in vivo* therapeutic use" (Final Office Action at page 21). Methods to treat patients (i.e., "*in vivo* therapeutic use") with the recited chimeric antibodies are not recited in the claims. The claims at issue recite chimeric antibodies which specifically bind to polypeptides of SEQ ID NO:1. The recited chimeric antibodies can be used, for example, to detect and/or purify polypeptides which are specifically bound by the recited antibodies. Therefore, the claimed chimeric antibodies are fully enabled, and no guidance "for *in vivo* therapeutic use" is necessary.

As set forth in *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971):

> The first paragraph of § 112 requires nothing more than objective enablement. How such a teaching is set forth, either by the use of illustrative examples or by broad terminology, is of no importance.
>
> As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.

Contrary to the standard set forth in *Marzocchi*, the Examiner has failed to provide any **reasons** why one would doubt that the guidance provided by the present specification would enable one to make and use the claimed antibodies which specifically bind to the recited polypeptides and variants of SEQ ID NO:1. Hence, a *prima facie* case for non-enablement has not been established with respect to the claimed antibodies which specifically bind to the recited polypeptides and variants of SEQ ID NO:1.

For at least the above reasons, reversal of this rejection is requested.

(9) CONCLUSION

The obviousness, written description and enablement rejections should be reversed, based on at least the arguments presented above.

If the USPTO determines that any additional fees are due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

This brief is enclosed in triplicate.

Respectfully submitted,

INCYTE CORPORATION

Date: 08 September 2003

Richard C. Ekstrom
Reg. No. 37,027
Direct Dial Telephone: (650) 843-7352

Date: 08 September 2003

Michelle M. Stempien
Reg. No. 41,327
Direct Dial Telephone (650) 843-7219

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Attachments

1. CLUSTALW alignment of FBP17 (g1255033) vs SEQ ID NO:1
2. Brenner et al., Proc. Natl. Acad. Sci. U.S.A. 95:6073-78 (1998)

APPENDIX - CLAIMS ON APPEAL

8. An isolated antibody selected from the group consisting of:

a) an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1 or SEQ ID NO:3, wherein the antibody binds to an epitope of a polypeptide of SEQ ID NO:1 or SEQ ID NO:3, and

b) an antibody which specifically binds to a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1 or SEQ ID NO:3, wherein the antibody binds to an epitope of a polypeptide having a naturally occurring amino acid sequence at least 90% identical to SEQ ID NO:1 or SEQ ID NO:3, said naturally occurring amino acid sequence having HS3C activity.

45. The antibody of claim 8, wherein the antibody is:

a) a chimeric antibody,

b) a single chain antibody,

c) a Fab fragment,

d) a $F(ab')_2$ fragment, or

e) a humanized antibody.

46. A composition comprising an antibody of claim 8 and an acceptable excipient.

48. The composition of claim 46, wherein the antibody is labeled.

50. A method of preparing a polyclonal antibody with the specificity of the antibody of claim 8, the method comprising:

a) immunizing an animal with a polypeptide comprising the amino acid sequence of SEQ ID NO:1, or an immunogenic fragment thereof, under conditions to elicit an antibody response,

b) isolating antibodies from said animal, and

c) screening the isolated antibodies with the polypeptide, thereby identifying a polyclonal antibody which binds specifically to a polypeptide comprising the amino acid sequence of SEQ ID NO:1.

51. An antibody produced by a method of claim 50.

52. A composition comprising the antibody of claim 51 and a suitable carrier.

53. A method of making a monoclonal antibody with the specificity of the antibody of claim 8, the method comprising:

a) immunizing an animal with a polypeptide comprising the amino acid sequence of SEQ ID NO:1, or an immunogenic fragment thereof, under conditions to elicit an antibody response,

b) isolating antibody producing cells from the animal,

c) fusing the antibody producing cells with immortalized cells to form monoclonal antibody-producing hybridoma cells,

d) culturing the hybridoma cells, and

e) isolating from the culture monoclonal antibody which binds specifically to a polypeptide comprising the amino acid sequence of SEQ ID NO:1.

54. A monoclonal antibody produced by a method of claim 53.

55. A composition comprising the antibody of claim 54 and a suitable carrier.

56. The antibody of claim 8, wherein the antibody is produced by screening a Fab expression library.

57. The antibody of claim 8, wherein the antibody is produced by screening a recombinant immunoglobulin library.

SeqServer®
biology *in silico*

ClustalW Results

Sequences | Help

Retrieval | BLAST2 | FASTA | ClustalW | GCG Assembly | Phrap | Translation

- [ ] 865744
- [ ] FBP

CLUSTAL W (1.7) Multiple Sequence Alignments

```
Sequence format is Pearson
Sequence 1: 865744          265 aa
Sequence 2: FBP             237 aa
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  86
Start of Multiple Alignment
There are 1 groups
Aligning...
Group 1: Sequences:   2      Score:2948
Alignment Score 1249
CLUSTAL-Alignment file created  [baaDSaiuV.aln]
CLUSTAL W (1.7) multiple sequence alignment


865744          MKRTVSDNSLSNSRGEGKPDLKFGGKSKGKLWPFIKKNKGATPEDFSNLPPEQRRKKLQQ
FBP             ----------------------------KIHCFRSLKRG-GVTPEDFSNFPPEQRRKKLQQ
                                           * : :  :*:. *.*******:***********

865744          KVDELNKEIQKEMDQRDAITKMKDVYLKNPQMGDPASLDHKLAEVSQNIEKLRVETQKFE
FBP             KVDDLNREIQKETDQRDAITKMKDVYLKNPQMGDPASLDQKLTEVTQNIEKLRLEAQKFE
                ***:**:***** ***************************:**:**:*******:*:****

865744          AWLAEVEGRLPARNEQARRQSGLYDSQNPPTVNNCAQDRESPDGSYTEEQSQESEMKVLA
FBP             AWLAEVEGRLPARSEQARRQSGLYDGQTHQTVTNCAQDRESPDGSYTEEQSQESEHKVLA
                *************.***********.*.  **.********************** ****

865744          TDFDDEFDDEEPLPAIGTCKALYTFEGQNEGTISVVEGETLYVIEEDKGDGWTRIRRNED
FBP             PDFDDEFDDEEPLPAIGTCKALYTFEGQNEGTISVVEGETLSVIEEDKGDGWTRIRRNED
                .****************************************** *****************

865744          EEGYVPTSYVEVCLDKNAKGAKTYI
FBP             EEGYFPTSYVEVYLDKNAKGAKTYI
                ****.******* ************
```

1: AAC52479. FBP 17...[gi:1255033]

BLink, Domains, Links

```
LOCUS       AAC52479                 237 aa            linear   ROD 04-APR-1996
DEFINITION  FBP 17.
ACCESSION   AAC52479
VERSION     AAC52479.1  GI:1255033
DBSOURCE    locus MMU40751 accession U40751.1
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (residues 1 to 237)
  AUTHORS   Chan,D.C., Bedford,M.T. and Leder,P.
  TITLE     Formin binding proteins bear WWP/WW domains that bind proline-rich
            peptides and functionally resemble SH3 domains
  JOURNAL   EMBO J. 15 (5), 1045-1054 (1996)
  MEDLINE   96183189
   PUBMED   8605874
REFERENCE   2  (residues 1 to 237)
  AUTHORS   Chan,D.C., Bedford,M.T. and Leder,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-NOV-1995) David C. Chan, Genetics, Harvard Medical
            School, 200 Longwood Avenue, Boston, MA 02115, USA
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..237
                     /organism="Mus musculus"
                     /strain="FVB"
                     /db_xref="taxon:10090"
     Protein         1..237
                     /product="FBP 17"
     CDS             1..237
                     /coded_by="U40751.1:<1..716"
                     /note="formin binding protein 17; contains SH3 domain"
ORIGIN
        1 kihcfrslkr ggvtpedfsn fppeqrrkkl qqkvddlnre iqketdqrda itkmkdvylk
       61 npqmgdpasl dqkltevtqn ieklrleaqk feawlaeveg rlparseqar rqsglydgqt
      121 hqtvtncaqd respdgsyte eqsqesehkv lapdfddefd deeplpaigt ckalytfegq
      181 negtisvveg etlsvieedk gdgwtrirrn edeegyfpts yvevyldkna kgaktyi
//
```

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "lnscaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

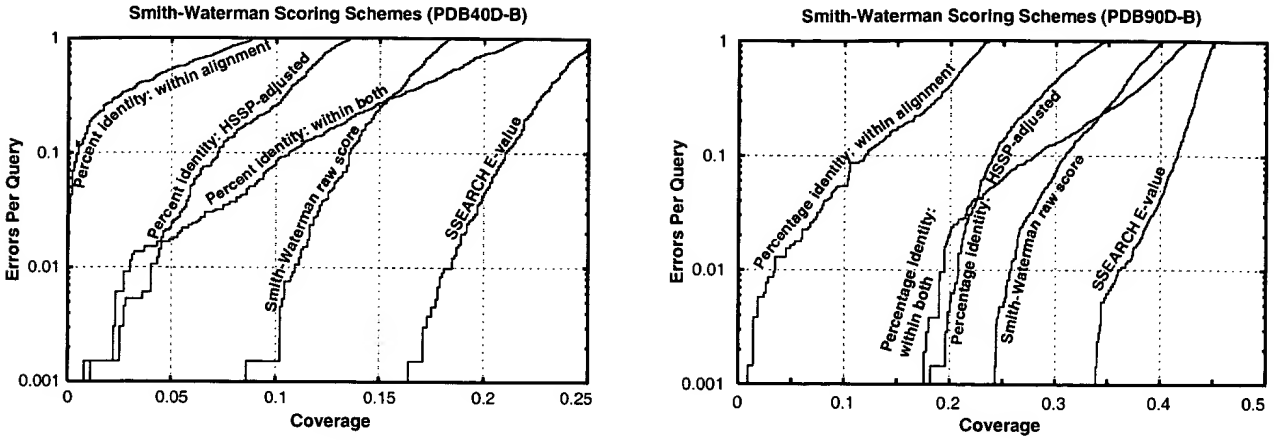


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

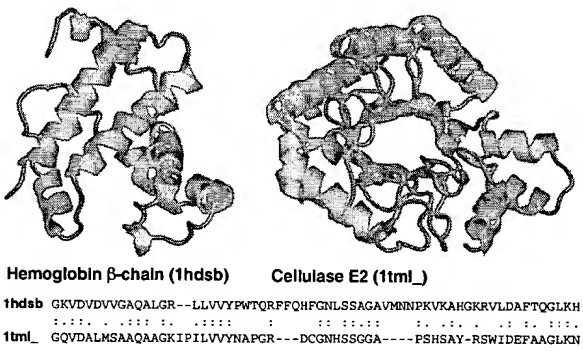


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

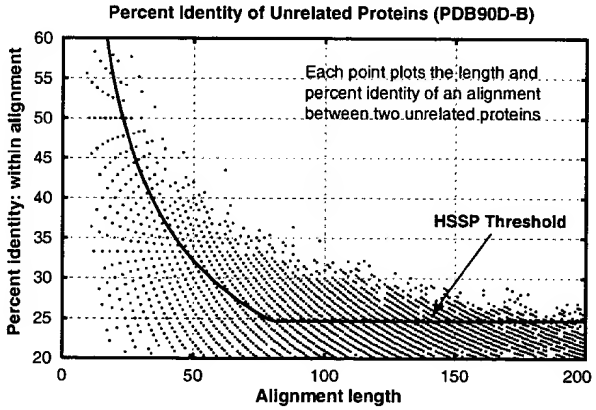


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

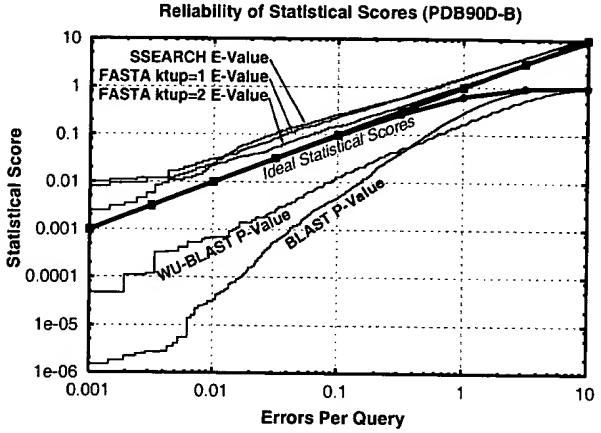


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

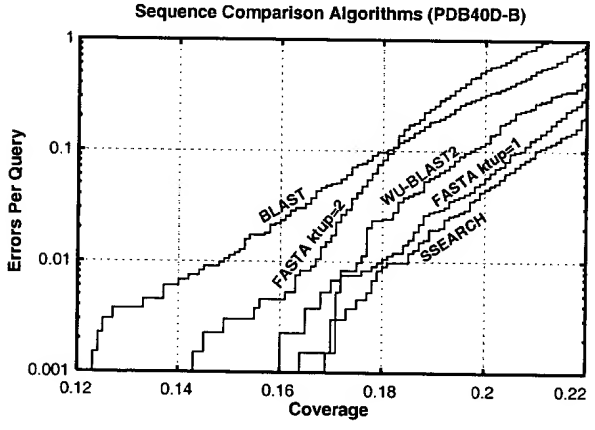
Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most



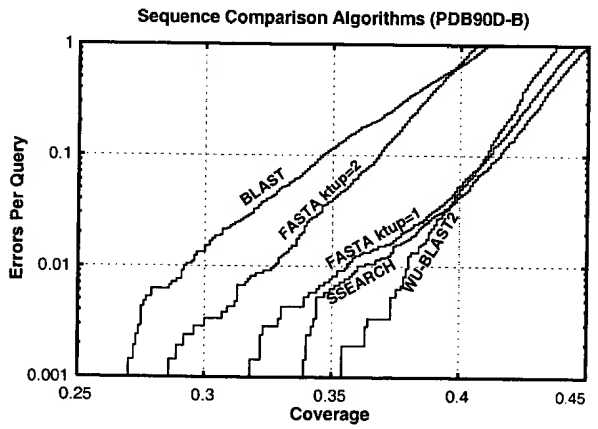


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

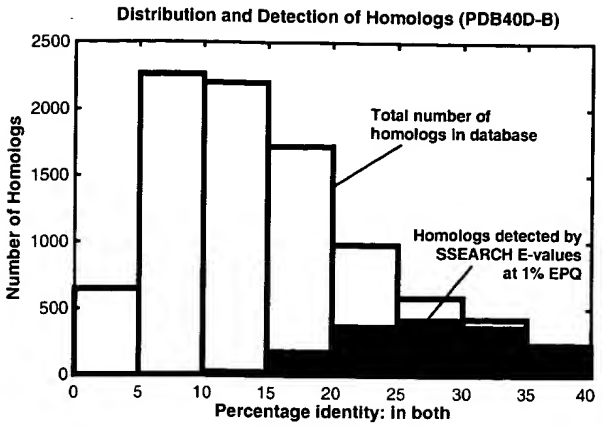


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.